NUKAT i Federacja Bibliotek Cyfrowych – pierwsze wyniki działań w kierunku integracji metadanych

Cezary Mazurek, Marcin Mielnicki, Krzysztof Sielski, Marcin Werla
Poznańskie Centrum Superkomputerowo-Sieciowe
{mazurek,marcinm,sielski,mwerla}@man.poznan.pl

1. Wstęp

Od początku obecnego stulecia obserwujemy w Polsce coraz bardziej dynamiczny rozwój naukowych i kulturowych zasobów informacyjnych dostępnych on-line. Projekty takie jak Wielkopolska (2002), Dolnośląska (2004) czy Kujawsko-Pomorska Biblioteka Cyfrowa (2005) wytyczały ścieżki organizacji regionalnych konsorcjów skupionych dookoła działań digitalizacyjnych [1]. Początkowo praktycznie całość tych działań miała charakter oddolny, często nie oparty na dedykowanym finansowaniu. Polska Biblioteka Internetowa (2003), pierwszy projekt rządowy związany z digitalizacją, posiadający duże i trwałe na przestrzeni kilku lat finansowanie okazał się ostatecznie porażką i jest obecnie podtrzymywany przy życiu przez Bibliotekę Narodową, do czasu przeniesienia zasobów do Cyfrowej Biblioteki Narodowej POLONA (2006). Na szczęście to jedyny negatywny przykład tego typu działań z ostatnich lat.

Duże zainteresowanie czytelników, jakim cieszyły się wspomniane powyżej inicjatywy oraz rosnąca stopniowo dostępność krajowego i unijnego finansowania, owocowały stale rosnącą liczbą bibliotek cyfrowych. W 2007 roku tych bibliotek było już kilkanaście i dawały dostęp do około 80 000 różnorodnych obiektów. Wtedy też w Poznańskim Centrum Superkomputerowo-Sieciowym powstała Federacja Bibliotek Cyfrowych, której głównym celem było ułatwienie dostępu do rozproszonych zasobów i zwiększenie ich wykorzystania, a sposobem realizacji tego celu była agregacja metadanych z poszczególnych bibliotek i udostępnianie ich poprzez jeden wspólny portal, wraz z odnośnikami do poszczególnych obiektów [2]. W ciągu 5 lat działalności FBC dalszy rozwój polskich bibliotek cyfrowych zaowocował około setką tego typu serwisów dostępnych on-line na początku 2013 roku, dających możliwość zapoznania się z ponad 1.2 mln obiektów ze zbiorów kilkuset instytucji kultury i nauki. W dużej mierze przyczyniły się do tego duże projekty takie jak Jagiellońska Biblioteka Cyfrowa, e-Biblioteka Uniwersytetu Warszawskiego czy Repozytorium Cyfrowe Instytutów Naukowych [3].

Równolegle do zasobów cyfrowych rozwijały się w Polsce w wielu dziedzinach bazy bibliograficzne (np. BazTECH czy Polska Bibliografia Literacka). Coraz więcej bibliotek udostępniało swoje katalogi on-line, a inicjatywy takie jak NUKAT czy KaRo ułatwiały dostęp do informacji w nich zawartych oraz w znaczący sposób wspierały dalszy rozwój m.in. poprzez możliwość wspólnego, skoordynowanego katalogowania (NUKAT) czy łatwego wyszukiwania i importu już istniejących opisów (KaRo). Coraz

częściej też bazy katalogowe czy bibliograficzne gromadziły odnośniki do cyfrowych postaci opisywanych publikacji dostępnych w bibliotekach cyfrowych. Biblioteki cyfrowe z kolei ułatwiały użytkownikom przejście od cyfrowej postaci obiektu i uproszczonych metadanych do pełnego opisu dostępnego on-line w katalogu bibliotecznym.

W 2010 roku dzięki finansowaniu Narodowego Centrum Badań i Rozwoju uruchomiony został projekt SYNAT, którego celem było „utworzenie uniwersalnej, otwartej, repozytoryjnej platformy hostingowej i komunikacyjnej dla sieciowych zasobów wiedzy dla nauki, edukacji i otwartego społeczeństwa wiedzy". Koordynatorem projektu został ICM UW, a wśród konsorcjum 16 instytucji jako jeden z głównych wykonawców znalazł się również PCSS[1]. Celem prac podejmowanych przez PCSS w ramach etapów A9 i A10 projektu SYNAT było m.in. opracowanie architektury systemu agregacji danych z rozproszonych, heterogenicznych systemów informacji naukowej oraz zbudowanie na podstawie tych informacji prototypowej bazy wiedzy obejmujące zarówno zasoby z bibliotek czy muzeów cyfrowych, jak i z muzealnych systemów inwentarzowych czy bibliotecznych systemów katalogowych.

Celem niniejszego artykułu jest poglądowe przedstawienie wyników prac PCSS, jakie udało się osiągnąć w zakresie agregacji i integracji danych z rozproszonych systemów informacji naukowej po 2.5 roku realizacji projektu SYNAT. Następny rozdział niniejszego tekstu omawia nowe podejście do agregacji danych z wielu źródeł wypracowane w ramach etapu A9 projektu SYNAT. Rozdział trzeci przedstawia podejście do reprezentacji wiedzy przy pomocy ontologii FRBRoo, a rozdział czwarty prezentuje wybrane aspekty zaimplementowanego procesu integracji i wzbogacania danych. Prototypowy interfejs użytkownika do powstałej w ten sposób bazy wiedzy opisano w rozdziale piątym. Rozdział szósty zawiera scenariusze wykorzystania bazy wiedzy m.in. w portalu Federacji Bibliotek Cyfrowych. Artykuł kończy podsumowanie i zarysowanie kierunków dalszych prac.

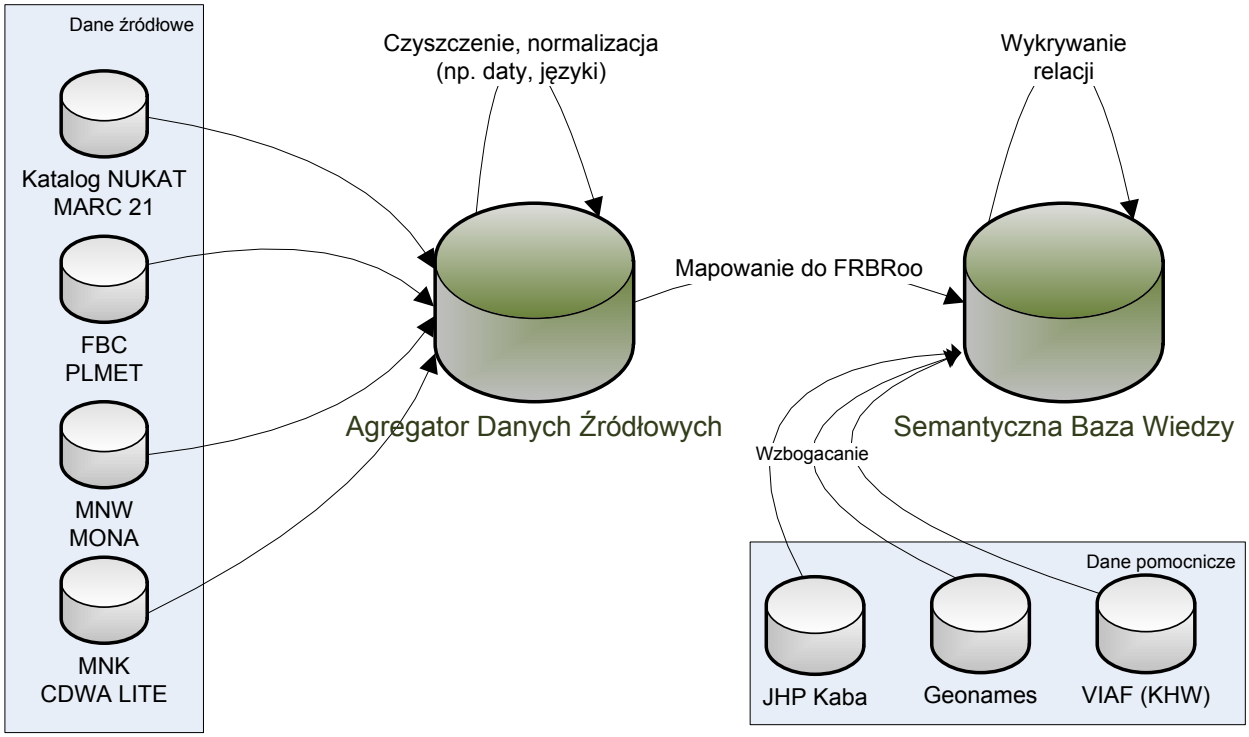
2. Agregacja danych z wielu źródeł

Zaimplementowany pierwotnie w Federacji Bibliotek Cyfrowych podstawowy mechanizm agregacji metadanych z polskich bibliotek cyfrowych opierał się na założeniu, że każde źródło danych powinno udostępniać swoje dane poprzez protokół OAI-PMH, w schemacie Dublin Core [4]. Schemat ten był przyjęty za model danych wspólny dla wszystkich bibliotek cyfrowych współpracujących z FBC i jeżeli któraś z bibliotek miała inny schemat metadanych, to po jej stronie leżał obowiązek opracowania i wykonania stosownej konwersji – mapowania danych. To założenie sprawdzało się w początkowym okresie i znacznie uprościło prace związane z implementacją FBC, jednak w dłuższej perspektywie okazało się niewystarczające. Po pierwsze wraz z powstawaniem kolejnych bibliotek cyfrowych okazało się, że schemat Dublin Core jest niewystarczający i wiele bibliotek rozszerza go o nowe elementy (takie jak chociażby „Miejsce wydania") [5]. Wskutek mapowania bogatszego schematu następuje niestety

[1] Opisywane w niniejszym artykule prace realizowane są w ramach projektu SYNAT finansowanego przez Narodowe Centrum Badań i Rozwoju (nr umowy: SP/I/1/77065/10)

utrata semantyki (znaczenia) danych (np. „Miejsce wydania” trafia do pola „Wydawca”) lub też utrata danych w całości, gdy rozszerzenia schematu nie są (celowo lub przypadkiem) uwzględniane w mapowaniu. Ponadto wymaganie dotyczące posiadania interfejsu OAI-PMH okazało się również problematyczne dla mniejszych instytucji.

Przy projektowaniu w ramach projektu SYNAT nowego mechanizmu integracji i agregacji danych o nazwie CLEPSYDRA (http://fbc.pionier.net.pl/pro/clepsydra) zmieniono te założenia. Przyjęto, że system agregacji powinien być w stanie pobrać dane z dowolnego systemu dostępnego on-line, oraz że pobierane powinny być dane w możliwie najbogatszej postaci – zarówno jeżeli chodzi o ilość danych, jak i o ich semantykę. Konwersja czy jakiekolwiek inne przetwarzanie danych powinno następować na późniejszym etapie, z uwzględnieniem specyfiki i oczekiwań aplikacji czy systemu, który będzie chciał zagregowane dane wykorzystać [6].



Rysunek 1. Schemat systemu agregacji i wzbogacania danych oraz konstrukcji bazy wiedzy.

Schematycznie przedstawiono to na Rysunku 1, gdzie widać dwa główne komponenty: Agregator Danych Źródłowych oraz Semantyczną Bazę Wiedzy. Rolą agregatora jest zebranie danych z różnych źródeł oraz ich czyszczenie i normalizacja. Komunikacja ze źródłami odbywa się poprzez system agentów, czyli małych programów, które są dedykowane do komunikacji z poszczególnymi klasami systemów informacji naukowej. Takie agenty potrafią wydobyć odpowiednie dane z konkretnego systemu (np. biblioteki cyfrowej czy systemu katalogowego) i zapisać je w odpowiedniej usłudze Agregatora Danych Źródłowych. Następnie, na podstawie predefiniowanych reguł przetwarzania realizowana jest konwersja (np. z formatu MARC binarny na MARC-XML czy z MARC-XML na PLMET) oraz czyszczenie i

normalizacja danych (np. ujednolicenie sposobu zapisu dat czy języków). W ramach projektu SYNAT opracowane zostały agenty wspierające źródła danych zgodne z protokołami OAI-PMH i OAI-ORE oraz źródła które udostępniają dane w postaci plików CSV. Dodatkowo przygotowano dedykowane agenty dla systemów NUKAT i Muzeum Narodowego w Warszawie.

Semantyczna Baza Wiedzy to system, który korzysta z Agregatora Danych Źródłowych – zebrane dane są okresowo pobierane i przetwarzane do postaci bazy wiedzy. W celu lepszej integracji zebranych danych źródłowych wykorzystywane są również źródła pomocnicze takie jak JHP Kaba, Geonames, TERYT czy VIAF. Baza wiedzy reprezentowana jest w ontologii FRBRoo, która została pokrótce przedstawiona w następnym rozdziale.

3. Reprezentacja wiedzy w ontologii FRBRoo

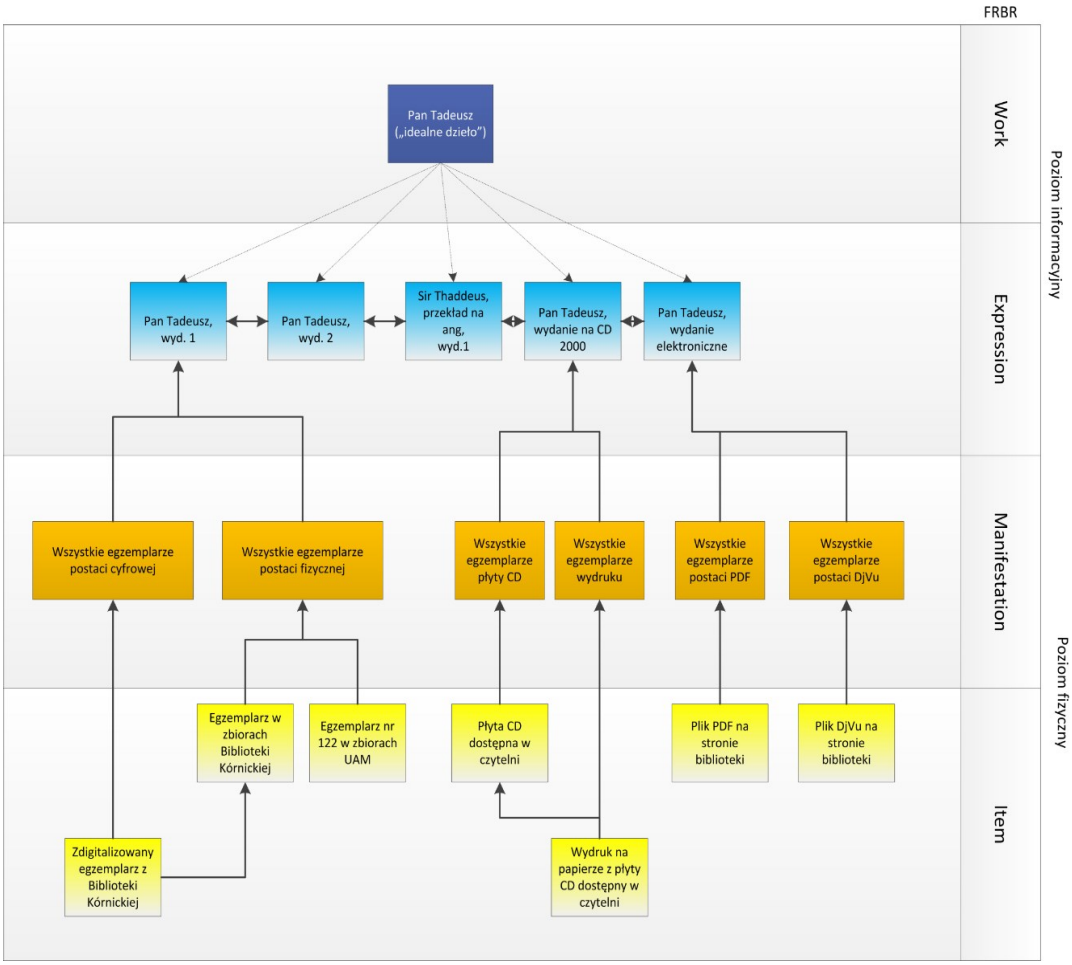
Pierwotnie do reprezentacji wiedzy w tworzonej przez PCSS w ramach projektu SYNAT bazie wiedzy wybrano ontologię CIDOC CRM [7]. Ontologia ta to zbiór konceptów i relacji miedzy nimi który umożliwia formalną reprezentację wiedzy na temat dziedzictwa kulturowego (w szczególności obiektów muzealnych). CIDOC CRM definiuje 86 różnych konceptów, takich jak na przykład nośnik informacji, osoba, temat czy język. Do tego wyspecyfikowano 139 relacji między nimi, np.: ma temat, brał udział w zdarzeniu, jest powiązany z.

FRBRoo [8] to rozszerzenie ontologii CIDOC CRM o koncepty z FRBR [9]. Rozszerzenie to wprowadza 33 nowe koncepty oraz 39 nowych relacji, pozwalając na lepsze wyrażenie wiedzy na temat obiektów bibliotecznych. Na Rysunku 2 widać przykładowe rozmieszczenie grup informacji na temat różnych form dzieła zatytułowanego „Pan Tadeusz” na czterech poziomach FRBR.

Najwyższy poziom to Dzieło (ang. Work). Pozwala on na wyrażenie cech pewnego ogólnego wytworu intelektualnego (w tym również artystycznego), bez przywiązywania się do konkretnej ustalonej formy intelektualnej tego dzieła. W uproszczeniu można to rozumieć jako pewne dzieło idealne funkcjonujące w umyśle autora. W momencie gdy dzieło to zostanie ustalone w konkretnej formie intelektualnej – np. autor napisze swoją książkę,spisze jej konkretny tekst – mamy do czynienia z drugim poziomem, czyli Realizacja (ang. Expression). Jest to jednak nadal poziom informacyjny – abstrahujemy na tym etapie od tego przy pomocy jakiego medium treść informacyjna została ustalona. Przykładami dwóch różnych Realizacji tego samego Dzieła mogą być np. tekst w oryginale i jego przekład na język obcy czy różniące się między sobą teksty dwóch kolejnych wydań tej samej książki. W tym drugim przypadku, jeżeli różnice pomiędzy tekstami wydań będą zbyt duże (np. zmiany tekstu całych rozdziałów), możemy mieć sytuację w której będą to ostatecznie dwa różne Dzieła. Trudno niestety podać tutaj bardzo jasne kryteria rozróżnienia.

Kolejny poziom FRBR to Materializacja (ang. Manifestation). Jest to poziom dotyczący fizycznego urzeczywistnienia realizacji dzieła i służy do wyrażenia cech wspólnych wszystkich egzemplarzy danego wydania. Nawiązując do przykładu z tłumaczeniem książki – jeżeli tekst tego tłumaczenia w identycznym brzmieniu zostanie wydany przez dwóch wydawców (lub jednego wydawcę na różnych mediach –

drukiem i jako e-booka), to będą to dwie Materializacje tej samej Realizacji tego samego Dzieła. Każda z tych Materializacji będzie miała przypisane cechy takie jak właśnie wydawca, rok wydania czy nośnik, jednak będą to tylko informacje ogólne takie jak np. format książki czy rodzaj okładki. Do wyrażenia cech specyficznych dla poszczególnych egzemplarzy (np. informacje o zniszczeniach, autografach czy proweniencji) służyć ma ostatni z poziomów FRBR czyli Egzemplarz (ang. Item).



Rysunek 2. Przykład umiejscowienia różnych postaci dzieła w modelu FRBR.

Opisany powyżej model jest modelem abstrakcyjnym i ogólnym. W zależności od charakteru przedmiotów i typu mediów, poszczególne poziomy mogą się w praktyce łączyć czy też tracić sens istnienia. Przykładem może być tutaj takie dzieło, które występuję tylko w jednym fizycznym egzemplarzu – unikalny rękopis wiersza czy pocztówka z wakacji. W takiej sytuacji Egzemplarz jest tylko jeden i możliwość rozróżnienia cech unikalnych poszczególnych egzemplarzy traci na znaczeniu. Inny przykład to Manifestacje zapisane na nośnikach cyfrowych. Tutaj co prawda cyfrowych kopii – Egzemplarzy – może być wiele, ale właściwie powinno być tak, że każda z nich jest identyczna z dokładnością do pojedynczego bitu. Znów więc możliwość wyrażenia różnić pomiędzy poszczególnymi Egzemplarzami może tracić na znaczeniu.

4. Integracja i wzbogacanie danych

Jak wspomniano w rozdziale 2, opisywany tu proces budowy bazy wiedzy w ogólności polega na mapowaniu zagregowanych i znormalizowanych danych do postaci ontologii FRBRoo, z wykorzystaniem pomocniczych źródeł danych, takich jak kartoteki i bazy lokalizacji geograficznych, osób, instytucji czy haseł przedmiotowych.

Proces ten nazywany procesem integracji i wzbogacania danych realizowany jest okresowo, w sposób w pełni zautomatyzowany. Ze względu na dużą ilość danych wejściowych (kilka milionów rekordów) i jeszcze większą ilość danych wynikowych (kilkaset milionów faktów/trójek w bazie wiedzy) nie ma możliwości ręcznej czy szczegółowo nadzorowanej przez człowieka realizacji takiego przetwarzania. W celu zapewnienia jakości przetwarzania dla każdego ze schematów danych wejściowych opracowane zostały unikalne reguły mapowania, a dodatkowo wynikowa baza wiedzy poddawana jest testom weryfikującym jej spójność. Przetwarzanie rekordów metadanych realizowane jest przy pomocy autorskiego narzędzia jMet2Ont (http://fbc.pionier.net.pl/pro/jmet2ont) [10].

Na Rysunku 3 przedstawiono wyrażony w XML-u fragment jednego z rekordów metadanych zagregowanego z biblioteki cyfrowej i przetworzonego do postaci PLMET. Jak widać. rekord ten składa się z pól ze schematu Dublin Core (pola poprzedzone przedrostkiem ‘dc:’) oraz pól specyficznych dla schematu PLMET[2] (pola poprzedzone przedrostkiem ‘plmet:’).

```
<plmet:metadata>
   <dc:title>Figliki albo rozlicznych ludzi przypadki dworskie [...]</dc:title>
   <dc:creator>Rej, Mikołaj (1505-1569)</dc:creator>
   <dc:contributor>Pencz, Georg (ca 1500-1550). Il.</dc:contributor>
   <dc:description>Dzieło pierwotnie współwydane z dziełem: Zwierziniec W ktorym
      rozmaitych [...]</dc:description>
   <dc:publisher>Drukarnia Macieja Wirzbięty</dc:publisher>
   <plmet:placeOfPublishing>Kraków</plmet:placeOfPublishing>
   <dc:date>1574</dc:date>
   <dc:language>pol</dc:language>
   <dc:coverage>16 w.</dc:coverage>
   <dc:subject>starodruki 16 w.</dc:subject>
   <dc:type>starodruk</dc:type>
   <dc:format>image/vnd.djvu</dc:format>
   <plmet:locationOfPhysicalObject>Biblioteka
      Jagiellońska</plmet:locationOfPhysicalObject>
   <dc:rights>Domena publiczna (public domain)</dc:rights>
   <plmet:digitisationSponsor>EFRR POIiŚ 11.1</plmet:digitisationSponsor>
 </plmet:metadata>
```

Rysunek 3. Fragment metadanych z biblioteki cyfrowej po agregacji i mapowaniu do schematu PLMET.

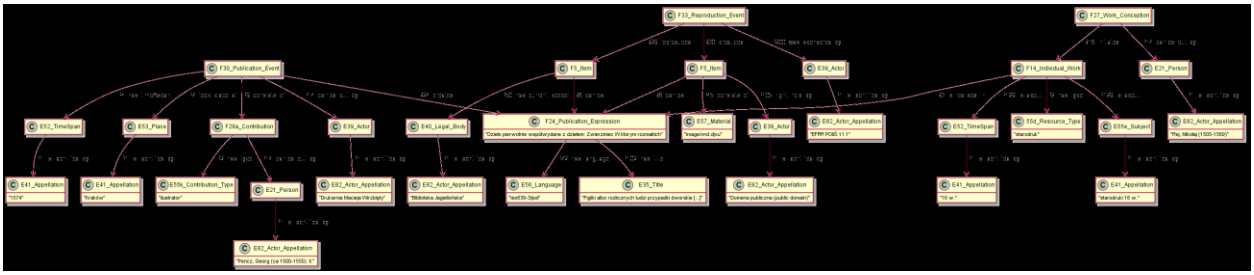
Rozszerzenie schematu Dublin Core umożliwiło w tym przypadku zachowanie informacji o tym, jakie jest miejsce wydania obiektu (plmet:placeOfPublishing), gdzie znajduje się obiekt fizyczny, który został

[2] Zob. https://dl.psnc.pl/community/x/UQDC

poddany digitalizacji (plmet:locationOfPhysicalObject) i jakie było źródło finansowania tej digitalizacji (plmet:digitisationSponsor).

Z kolei na Rysunku 4 przedstawiono grafową wizualizację powyższego rekordu po przetworzeniu go do postaci bazy wiedzy. W efekcie tej operacji uzyskano 33 powiązane ze sobą obiekty opisane przez 78 trójek RDF. Na etapie wzbogacania danych udało się ustalić powiązania informacji ze źródłowego rekordu z pomocniczymi bazami danych (w nawiasach podano oznaczenia klas z ontologii FRBRoo):

- Zapis *Rej, Mikołaj (1505-1569)*został rozpoznany jako informacja o osobie (klasa E21 Person) i powiązany z pozycją z VIAF http://viaf.org/viaf/61585459
- Zapis *Pencz, Georg (ca 1500-1550). Il.* został rozpoznany jako informacja o osobie (klasa E21 Person) i powiązany z pozycją z VIAF http://viaf.org/viaf/64120782
- Zapis *Kraków* został rozpoznany jako informacja o miejscu (klasa E53 Place) i powiązany z pozycją z Geonames http://www.geonames.org/3094802
- Zapis *pol* został rozpoznany jako informacja o języku (klasa E56 Language) i powiązany z pozycją z Lexvo http://lexvo.org/id/iso639-3/pol
- Zapis *starodruki 16 w.* został rozpoznany jako informacja o haśle przedmiotowym (klasa E55h Subject Hierarchy) i powiązany z hasłem z NUKAT s 2010216717 (Stare druki -- 16 w.)
- Zapis *Biblioteka Jagiellońska* został rozpoznany jako informacja o instytucji (klasa E40 Legal Body) i powiązany z hasłem w VIAF http://viaf.org/viaf/148485690



Rysunek 4. Graf wizualizujący wynik mapowania rekordu metadanych z Rysunku 3 do ontologii FRBRoo.

Powyższy przykład pokazuje, że wypracowana metodyka bazy wiedzy daje wymierne efekty, a wykorzystanie pomocniczych źródeł danych zwiększa integralność danych wynikowych – w powyższym przykładzie rekord z biblioteki cyfrowej został skojarzony m.in. z informacjami z NUKATu [11]. Oczywiście opisywane tu rozwiązanie nie jest w stanie poradzić sobie z dowolnymi danymi – im większy będzie stopień normalizacji danych wejściowych, tym większa szansa na wyższą jakość danych wynikowych. Dla niektórych elementów informacji źródłowych możliwe jest też podjęcie próby normalizacji wszystkich wartości, gdyż unikalnych wartości do przejrzenia i uwzględnienia w mapowaniu nie jest zbyt wiele (do kilkuset w przypadku pól „język” czy „prawa”) lub też wartości te cechują się stosunkowo dużą regularnością formy zapisu i mogą być przetwarzane automatycznie (pole „data”).

Dodatkowym efektem opisanego powyżej procesu przetwarzania danych źródłowych i pomocniczych do postaci bazy wiedzy może być poprawa jakości tych danych. Po pierwsze dostawcy danych źródłowych

mogą być zainteresowani dostępem czy zwrotnym transferem tych danych już po ich czyszczeniu i normalizacji. Po drugie, w przypadku źródeł danych pomocniczych o podobnym charakterze możliwe jest podjęcie próby wzbogacenia jednego z tych źródeł na podstawie informacji w drugim. Dla przykładu w kartotece KABA istnieją osobne hasła reprezentujące miejscowość Kcynia (s 97053818) i gminę Kcynia (s 2012307665). Nie ma jednak między nimi żadnego powiązania (tzw. tropu), które wskazywałoby zależność między gminą i leżącą na jej terenie miejscowością (jest to akurat zgodne z zasadami tworzenia kartoteki KABA). Obydwa z tych haseł można powiązać automatycznie z ich odpowiednikami w bazie Geonames (miasto: http://www.geonames.org/3096385/kcynia, gmina: http://www.geonames.org/7533422/kcynia), gdzie takie powiązanie już funkcjonuje. Automatyczne wykrycie różnic w powiązaniach pomiędzy dwiema różnymi bazami mówiącymi o tych samych obiektach (lokalizacjach) jest w tym przypadku możliwe, a wygenerowany na tej podstawie raport może zostać następnie użyty do wzbogacenia przeanalizowanych w ten sposób baz.

Kolejnym naturalnym etapem prac po zaprojektowaniu i implementacji mechanizmów automatycznej budowy bazy wiedzy było przygotowanie interfejsów dostępowych. Rozdział 5 przedstawia pokrótce prototypowy interfejs dostępowy do bazy wiedzy, a w rozdziale 6 zaprezentowane są wybrane scenariusze wykorzystania tej bazy m.in. w interfejsie użytkownika Federacji Bibliotek Cyfrowych.

5. Prototypowy interfejs do bazy wiedzy

Semantyczna baza wiedzy, w odróżnieniu do systemów opartych na relacyjnych bazach danych, operuje na grafowym modelu informacji. Oznacza to, że w bazie semantycznej nie mamy do czynienia z wielokolumnowymi tabelami zawierającymi kompleksowe informacje na temat danego rodzaju obiektów (lub ewentualnie relacjami do innych tabel). Zamiast tego są pojedyncze obiekty powiązane pomiędzy sobą relacjami. Prezentacja takich danych może być zwizualizowana w postaci grafu, jednak w przypadku rozległej bazy semantycznej problemem jest to, jaką część grafu prezentować jednorazowo. Dla przykładu, jeżeli użytkownik zapyta się o informacje dotyczące autora, to czy należy zaprezentować wyłącznie jego imię i nazwisko, czy może jeszcze daty narodzin i śmierci, miejsce urodzenia i dzieła? A jeżeli pokazujemy miejsce urodzenia, to czy od razu podać też nazwiska innych urodzonych tam twórców? Czy przy dziełach podać ich wydawców? Czy podać słowa kluczowe? Czy przy tych słowach kluczowych podać inne dzieła na ten sam temat lub podobne słowa kluczowe? Każda z tych decyzji wpływa na ilość jednorazowo prezentowanych informacji, co przekłada się na satysfakcję użytkownika (lub jej brak – zarówno w przypadku braku jak i nadmiaru informacji) oraz na obciążenie systemu informatycznego.

W przypadku prostych modeli danych łatwiej jest określić, które informacje prezentować od razu, a które powinny być dostępne po dodatkowym żądaniu. Ontologia FRBRoo nie jest jednak modelem prostym (119 konceptów, 178 relacji). Do tego jej natura skłania do przyjmowania różnych perspektyw prezentowania informacji – wiedzę o tym, że książka Mikołaja Reja została wydana w 1574 r. w Krakowie przez Drukarnię Macieja Wirzbięty można spojrzeć z perspektywy twórcy, wydawcy, miejsca

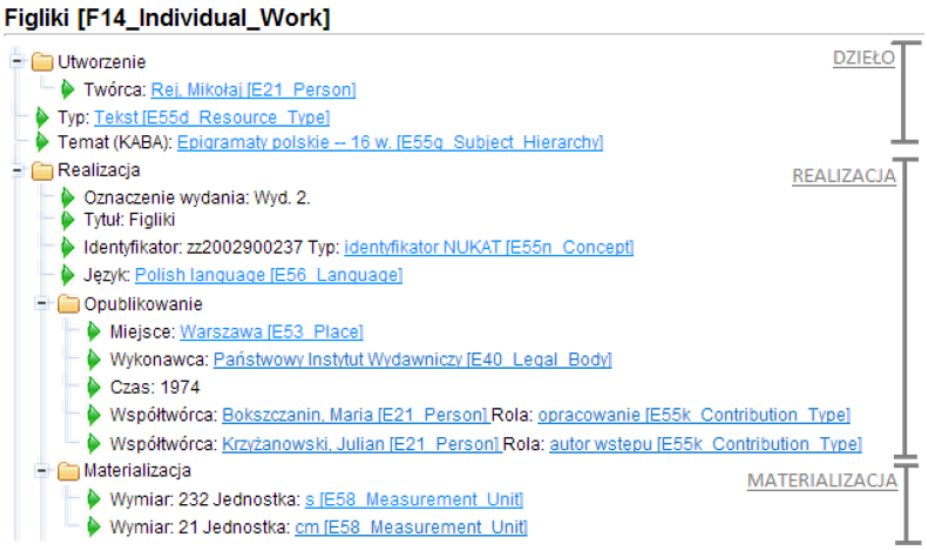
jak i roku wydania – wszystko zależy od tego, w którym miejscu widocznego na Rysunku 4 grafu użytkownik rozpocznie eksplorację bazy wiedzy.

W trakcie projektu SYNAT w PCSS wypracowane zostało rozwiązanie polegające na oznaczaniu krawędzi grafu (czyli relacji) specjalnymi znacznikami informującymi, czy informację jaką wnosi ta krawędź wyświetlić bezpośrednio, czy w postaci hiperłącza prowadzącego do dalszej części informacji, czy też ją pominąć. Na tej podstawie opracowane zostało uniwersalne narzędzie do prezentacji semantycznych baz wiedzy oznakowanych w opisany powyżej sposób. Przykładowy zrzut ekranu z tego narzędzia widoczny jest na Rysunku 5.

Jak widać, w tym przypadku system od razu wyświetlił m.in. informacje o tym kto jest twórcą książki, kto wydawcą oraz szczegółowe informacje na temat nośnika – egzemplarza publikacji. Nie prezentowano jednak od razu żadnych dodatkowych informacji na temat autora, pozwalając użytkownikowi kliknąć na odpowiednie hiperłącze i przejść do widoku bazy, gdzie ten autor, a nie jego publikacja, będzie podstawowym obiektem zainteresowania.

Na Rysunku 6 przedstawiono kolejny zrzut ekranu interfejsu bazy wiedzy, tym razem prezentujący informacje o wydaniu Figlików Mikołaja Reja, opublikowanym w Warszawie w 1974 roku. Na tym zrzucie ekranu zarysowuje się wyraźnie podział na dzieło, realizację i materializację z modelu FRBR, tak jak to opisano w rozdziale 3.

Rysunek 5. Przykładowy zrzut ekranu z prototypowego interfejsu dostępowego do bazy wiedzy.



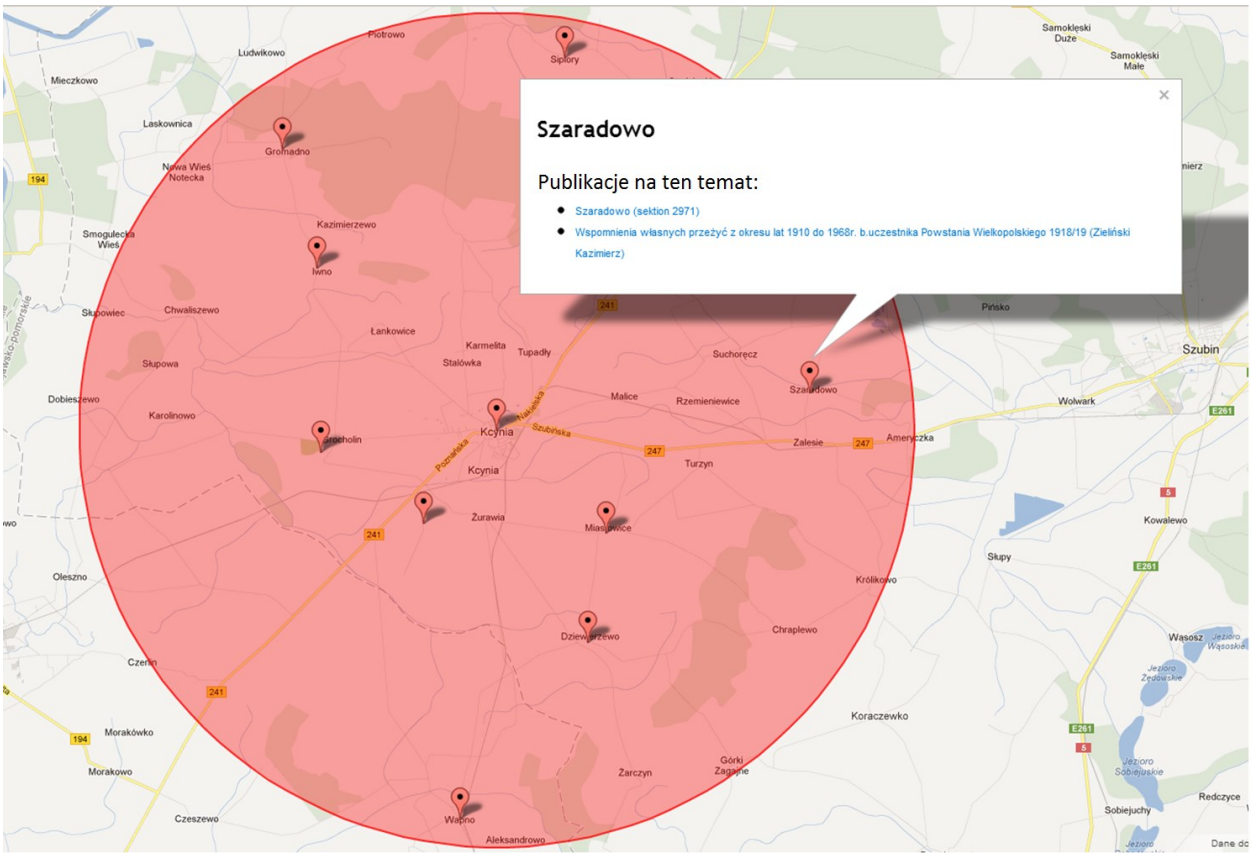
Rysunek 6. Wyróżnienie poziomów FRBR w danych przetworzonych do postaci ontologii FRBRoo i zaprezentowanych w prototypowym interfejsie dostępowym bazy wiedzy.

6. Scenariusze wykorzystania bazy wiedzy

Opisany w poprzednim rozdziale prototypowy interfejs dostępowy do bazy wiedzy to tylko jeden ze sposobów korzystania z tej bazy. Docelowo planowane jest stopniowe wzbogacanie funkcji oferowanych przez portal Federacji Bibliotek Cyfrowych o nowe możliwości oparte na technologiach semantycznych, informacjach zgromadzonych w bazie wiedzy i narzędziach opracowanych na potrzeby jej utworzenia.

Jednym z możliwych scenariuszy jest wprowadzenie wyszukiwania geograficznego. W takim scenariuszu, jeżeli zapytanie wprowadzone przez użytkownika zostanie rozpoznanie jako lokalizacja geograficzna, możliwe będzie wyświetlenie wyników związanych nie tylko z tą lokalizacją, ale również z lokalizacjami pobliskimi czy podrzędnymi lub też wręcz zaprezentowanie wyników na mapie. Przykład takiego działania przedstawiono na Rysunku 7.

Innym przykładem może być wykorzystanie hierarchii tematów. Gdy użytkownik w wyszukiwarkę wprowadzi dość ogólne hasło (np. fizyka), poza zaprezentowaniem mu wyników opisanych wprost tym hasłem, można mu też podpowiedzieć inne słowa węższe znaczeniowo (np. promieniotwórczość, mechanika, …) wraz z liczbą wyników które dodatkowo uzyska zmieniając w sugerowany sposób swoje zapytanie.



Rysunek 7. Przykład wyszukiwania geograficznego – użytkownik wyszukujący miejscowości Kcynia otrzymuje informacje o publikacjach powiązanych również z pobliskimi miejscowościami.

Scenariuszy tego typu można wskazać więcej. Możliwe jest m.in. sugerowanie innego sposobu zapisu nazwiska autora czy też użycie w wyszukiwaniu nie tylko oficjalnego nazwiska, ale i literackiego pseudonimu. Można również sugerować użytkownikowi obiekty powiązane w ciekawy sposób z obiektem aktualnie prezentowanym – np. przy wyświetlaniu danych na temat obrazu można wyświetlić listę publikacji dotyczących twórczości autora tego dzieła.

7. Podsumowanie

W niniejszej pracy przedstawiono realizowane przez PCSS w ramach projektu SYNAT podejście do agregacji, wzbogacania i integracji danych z heterogenicznych rozproszonych systemów informacji naukowej do postaci bazy wiedzy. Zrealizowany w trakcie projektu SYNAT prototyp z powodzeniem wykorzystany został do zebrania w jednej bazie łącznie kilku milionów rekordów z kilkudziesięciu źródeł danych – w tym bibliotek, muzeów i archiwów cyfrowych, katalogów bibliotecznych (w tym bazy NUKAT) oraz części danych z systemu inwentaryzacji zabytków Muzeum Narodowego w Warszawie. Do integracji danych wykorzystano ontologię FRBRoo oraz pomocnicze źródła danych takie jak Geonames, TERYT, VIAF czy słownik JHP KABA.

Poza opracowaniem prototypowego interfejsu dostępowego do bazy wiedzy, planuje się również stopniowe wzbogacanie funkcji portalu FBC o nowe możliwości, których implementacja możliwa będzie dzięki powstałej bazie wiedzy i towarzyszącym jej narzędziom.

Dalsze prace badawcze i rozwojowe skupią się na przyłączaniu nowych źródeł danych i opracowywaniu zasad reprezentacji uzyskanych w ten sposób danych w bazie wiedzy, a także na nowych wizualnych interfejsach eksploracji bazy wiedzy i wykorzystywania zawartych w niej informacji w pracy naukowców korzystających z narzędzi takich jak Wirtualne Laboratorium Transkrypcji (http:// wlt.synat.pcss.pl/). Kolejnym wyzwaniem będzie również włączanie informacji z semantycznych baz danych, które PCSS utrzymuje na potrzeby naukowych projektów humanistycznych, w ramach platformy e-humanistyki dostępnej pod adresem http://ehum.psnc.pl/.

Bibliografia

[1] C. Mazurek, M. Stroiński, M. Werla, *Wdrażanie regionalnych bibliotek cyfrowych w sieci PIONIER w oparciu o środowisko dLibra*, [w:] *INFOBAZY 2005 – Bazy Danych dla Nauki, Gdańsk, 25 – 27 wrzesień, 2005,* Gdańsk 2005

[2] C. Mazurek, M. Werla, *Federacja Bibliotek Cyfrowych – studium przypadku*, [w:] *Biblioteki cyfrowe*, red. M. Janiak, M. Krakowska, M. Próchnicka, Warszawa 2012

[3] D. Czarnocka-Cieciura, D. Gazicka-Wójtowicz, B. Górczyńska, K. Lis, *Realizacja projektu RCIN*, [online], http://www.petea.home.pl/apan/files/user_files2/iii%20spotkania_do%20publikacji/rcinapan2011_.pdf [dostęp: 26.04.2013]

[4] C. Mazurek, M. Stroiński, M. Werla, J. Węglarz, *Distributed Services and Metadata Flow in the Polish Federation of Digital,* [w:] *2011 International Conference on Information Society (i-Society)*, 2011

[5] A. Wróbel, J. Potęga, *The Dublin Core Metadata Element Set, Ver. 1.1 a potrzeby i oczekiwania bibliotekarzy cyfrowych – analiza przypadków*, [w:] *Polskie Biblioteki Cyfrowe 2009*, Poznań 2010

[6] C. Mazurek, M. Mielnicki, A. Nowak, M. Stroiński, M. Werla, J. Węglarz, *Architecture for Aggregation, Processing and Provisioning of Data from Heterogeneous Scientific Information Services*, [w:] *Intelligent Tools for Building a Scientific Information Platform: Advanced Architectures and Solutions*, red. R. Bembenik, Ł. Skonieczny, H. Rybiński, M. Kryszkiewicz, M. Niezgódka, Springer Berlin Heidelberg 2013

[7] *ICOM/CIDOC Documentation Standards Group and CIDOC CRM Special Interest Group, Definition of the CIDOC Conceptual Reference Model. Version 5.0.4*, [online], http://www.cidoc-crm.org/docs/cidoc_crm_version_5.0.4.pdf [dostęp: 26.04.2013]

[8] *International Working Group on FRBR and CIDOC CRM Harmonisation, FRBR object-oriented definition and mapping to FRBRer (Version 1.0.2)*, [online], http://www.cidoc-crm.org/docs/frbr_oo/frbr_docs/FRBRoo_V1.0.2.pdf [dostęp: 26.04.2013]

[9] IFLA Study Group on the Functional Requirements for Bibliographic Records, Functional Requirements for Bibliographic Records. Final Report, 1997, [online]. Dostępny w World Wide Web: http://www.ifla.org/en/publications/functional-requirements-for-bibliographic-records

[10] J. Walkowska, M. Werla, *Advanced Automatic Mapping from Flat or Hierarchical Metadata Schemas to a Semantic Web Ontology,* [w:] *Lecture Notes in Computer Science, vol. 7489*, Springer Berlin Heidelberg 2012

[11] C. Mazurek, K. Sielski, J. Walkowska, M. Werla, *From MARC21 and Dublin Core, through CIDOC CRM: First Tenuous Steps towards Representing Library Data in FRBRoo*, [w:] *CIDOC 2012*, 2012